HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Corpora in Text-Based Russian Studies

Kopotev, Mikhail

Palgrave Macmillan
2020

Kopotev , M , Mustajoki , A & Bonch-Osmolovskaya , A 2020 , Corpora in Text-Based Russian Studies . in D Gritsenko , M Wijermars & M Kopotev (eds) , The Palgrave Handbook of Digital Russia Studies . Palgrave Macmillan , Cham , pp. 299-317 . https://doi.org/10.1007/978-3-030-42855-6_17

http://hdl.handle.net/10138/326118
https://doi.org/10.1007/978-3-030-42855-6_17

cc_by
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Check for updates

CHAPTER 17

Corpora in Text-Based Russian Studies

Mikhail Kopotev, Arto Mustajoki, and Anastasia Bonch-Osmolovskaya

17.1 Introduction

This chapter focuses on textual data that are collected for a specific purpose, which are usually referred to as *corpora*. Scholars use corpora when they examine existing instances of a certain phenomenon or to conduct systematic quantitative analyses of occurrences, which in turn reflect habits, attitudes, opinions, or trends. For these contexts, it is extremely useful to combine different approaches. For example, a linguist might analyze the frequency of a certain buzzword, whereas a scholar in the political, cultural, or sociological sciences might attempt to explain the change in language usage from the data in question. This handbook is no exception: the reader will find several chapters (for additional information, see Chaps. 26, 23, 29 and 24) that are either primarily or secondarily based on Russian textual data.

Russian text-based studies represent a well-established area of science, unknown in part to Western readers due to the language barrier. However, this

M. Kopotev (✉)
Higher School of Economics (HSE University), Saint Petersburg, Russia
e-mail: mkopotev@hse.ru

A. Mustajoki
Higher School of Economics (HSE University), Saint Petersburg, Russia

University of Helsinki, Helsinki, Finland
e-mail: arto.mustajoki@helsinki.fi

A. Bonch-Osmolovskaya
Higher School of Economics (HSE), Moscow, Russia
e-mail: abonch@hse.ru

© The Author(s) 2021
D. Gritsenko et al. (eds.), *The Palgrave Handbook of Digital Russia Studies*, https://doi.org/10.1007/978-3-030-42855-6_17

should not overshadow the existence of well-developed tools and promising results (Dobrushina 2007; Mustajoki and Pussinen 2008; Plungian 2009). Naturally, scholars in linguistics have made the most visible progress in corpus studies, offering a wide spectrum of data (described in Sect. 17.3 of this chapter) and a range of corpus-based methods that are reflected in recent publications (Plungian 2009; Plungian and Shestakova 2014; Zabotkina 2015; Lyashevskaya 2016; Kopotev et al. 2018).

In the chapter, we describe existing textual resources in Russian, from available online sites to DIY ("do-it-yourself") corpora, with a special focus on two of the most significant examples: the Russian National Corpus and the Integrum database. Finally, in the last section, we present two cases of corpus-based analysis: the first investigates the collective mnemonic patterns for names of decades in Soviet and post-Soviet history and the second concerns political trends in modern Russia.

T. McEnery and A. Wilson (1996, 24) offer the following definition of a corpus:

> Corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a *finite-sized body* of *machine-readable* text, sampled in order to be *maximally representative* of the language variety under consideration. (Italics added)

Three features of this definition need to be highlighted as they constitute the quality criteria for any corpus data. The first is that it is finite-sized. This means that the number of tokens is known so the user can apply various statistics to the data, ranging from simple frequency rankings to sophisticated neuronal algorithms. The second quality is that it is in a machine-readable format that allows users to conduct quick searches within an unlimited amount of data, from Tolstoy's masterpieces to ordinary texts available on the internet. The third quality is maximal representativeness, which makes it possible to draw conclusions from a finite number of examples on the infinity of a language or its variety. In this sense, the usage of corpora in the humanities makes it similar to a hard science, meaning that the results are calculable and replicable, and thus able to be tested.

17.2 The Web as a Corpus

The emergence of search engines such as Yahoo, and later Google, has made it possible to explore the World Wide Web and its expanding massive number of sites. This development has given rise to new verbs such as "googling" (meaning to search on google.com) and "yandexing" (to search on yandex.ru). The Russian part of the global internet is often referred to as the Runet (for additional information, see Chap. 16). This includes not only sites under the country code's top-level domain. RU but every site

available in the Russian language. Runet had a six percent share of all internet sites for 2018, putting it in second place after advanced English (see Usage 2019). However, a clear differentiation should be made between search engines that index websites and corpora. Search engines that index websites allow users to make searches, whereas corpora constitute data, the results of which are controlled and replicable.

Whichever commercial search engine is used, it is primarily intended to deliver information that includes, first and foremost, marketing material that targets specific consumer groups. One can, of course, use the internet for information mining but the results may be scientifically unreliable without additional verification. Information from data mining tends to contain drivel attributable to varying spelling norms, scanning errors, fluctuation in internet communication, and so forth. As Adam Kilgarriff observes:

> [L]ike Borges's Library of Babel, [the internet] contains duplicates, near duplicates, documents pointing to duplicates that may not be there, and documents that claim to be duplicates but are not. (Kilgarriff 2001, 342)

A simple internet search yields *a priori* unknown results, which are usable only if they are task-specific and the researcher is cognizant of all the limitations. Even then, using the internet as a source is fraught with serious risks. Among the most serious is the fact that users do not control the data they search and they do not control the search engines they use (see Bozdag 2013; Flaxman et al. 2016).

It is difficult to conduct data-based research without texts that are *reliable* and *accessible*. By reliable, we mean texts that are consistently of high quality, and by accessible, we refer to texts that are easily obtainable. A general caveat with regard to the data that are available online is that the smaller the text and the more unique its contents, the more reliable the source should be. If the features of an individual text are not crucially important, then any potential noise in the data can be ignored, at least to some extent. A large amount of noisy data may nonetheless be used effectively to study general tendencies in the language variety under consideration. For example, a noise would be caused by errors related to a source, as in mixing Latin and Cyrillic letters after Optical Character Recognition (OCR) processing, and these are dissolved in the total mass of data.

Electronic texts that are available on the internet fall into one of three, uneven, categories: the majority are insufficiently prepared (e.g., a source is not reliable), error-filled (e.g., inaccurately digitized), or non-authorized (such as a doubtful copyright status). A smaller amount of textual data, with more attention given to their quality, can be further categorized as non-linguistic collections, or "electronic libraries," and linguistically oriented collections, or "linguistic corpora." Naturally, the distinction between non-linguistic and linguistic data is somewhat vague and depends heavily on the task at hand, the

main difference being whether or not the data are linguistically annotated, that is, enriched with linguistic information.

17.3 Electronic Libraries

Collections of texts are not corpora in the strict sense of the term. However, large text collections have a wide circulation in digital studies and are reliable resources for Russian studies. The largest of these collections on Runet are Moshkov's Library (www.lib.ru) and Librusec (www.lib.rus.ec).[1] Access to both sites is free and includes massive collections of fictional and non-fictional Russian texts. Furthermore, both could serve as good initial sources for big-data studies in Russian digital humanities (for more, see Chap. 29).

When the research objective is to analyze literary masterpieces, the sources need to be more carefully selected. In this context, Runet has three useful websites that aim to provide high-quality data. The first is the Fundamental Electronic Library of "Russian Literature and Folklore" (www.feb-web.ru), which is a fast-developing collection of belles lettres that follows the strict guidelines of academic publications, enriched with commentaries and an extended reference apparatus. The website contains fiction from the eighteenth to the twentieth century as well as Old Russian literature and folklore. The second resource is the Russian Virtual Library (www.rvb.ru). The content, principles, and developers of this collection partly overlap with the Fundamental Electronic Library, although the latter focuses more on published Russian texts from the eighteenth century, the *fin de siècle*, and from Soviet underground poetry. The third resource, lib.pushkinskijdom.ru, is maintained by the Institute of Russian Literature (RAS, also known as *Pushkinskij dom*). This site provides access to thousands of texts from the ninth to the twentieth century. These consist mainly of fiction and poetry, but also memoirs, critical reviews, and critical bibliographies. A true gem of the collection is the library of Old Russian literature, which includes most of the surviving ancient texts and their Russian translations.

17.4 Linguistic Corpora[2]

While the aforementioned sources are sufficient for many researchers, linguists require resources that are specifically designed for their analyses of language phenomena. These are referred to as "linguistic corpora," which means that the entries are enriched with specific linguistic information. Some examples of this are tokenization, lemmatization, part-of-speech tagging, and syntactic relations. This detailed information enables scholars who are more interested in the linguistic content of the texts to search in sources that are more directly oriented to linguistic information.

The dawn of computer-assisted research in the Russian language occurred at the turn of the twenty-first century, which was shortly after the emergence of

resources specifically designed to meet the needs of linguistics scholars, namely linguistic corpora. Russian corpus linguistics is currently a highly developed branch of linguistic studies and is well represented in national computational linguistic landscapes (see the "Dialogue" conferences at www.dialog-21.ru/en) and in international collaboration (see, e.g., Erjavec et al. 2010; Nivre et al. 2018). The following extensive "big data" resources were made available from the beginning, presented below in an ascending order of tokens:

- the Araneum Russicum corpora of 1.2 billion tokens (Benko 2014);
- the ruWac: the Russian portion of the project "The Web as a Corpus" of 1.3 billion tokens (Sharoff and Nivre 2011);
- the Taiga corpus of 5 billion tokens (Shavrina and Shapovalova 2017);
- ruTenTen of 14.5 billion tokens, a member of the commercial TenTen corpus family (Jakubíček et al. 2013);
- General Internet Corpus of Russian of 19.8 billion tokens (GICR; see Belikov et al. 2013).

The above list of corpora and resources is by no means comprehensive, and many smaller, more specific and more deeply annotated corpora are available for academic use (see the catalogue at www.ruscorpora.ru/new/corpora-other.html). There are also various historical and parallel corpora, as well as corpora that are not publicly available, which are beyond the scope of this chapter (see reviews in Mitrenina 2014; Mikhailov and Cooper 2016; Kopotev et al. 2018). Nonetheless, in many cases, the best available option is to create a task-specific corpus.

A do-it-yourself (DIY) corpus eliminates many issues caused by raw internet data, such as repetition, disproportion, and babelization (language mixture). Many special tools have been developed to create DIY corpora, typically referred to as a "concordancer" or "corpus manager" (see https://en.wikipedia.org/wiki/Corpus_manager). Researchers can use these programs to look up contexts, construct lists of keywords or frequencies, analyze word co-occurrences, and determine the distribution of words across texts or topics. A reliable option that is available to scholars is the commercial Sketch Engine service and its non-commercial version, No Sketch Engine (www.sketchengine.eu/nosketch-engine). The service includes many specific linguistic tools that are available upon registration.

17.4.1 The Russian National Corpus (www.ruscorpora.ru)

A national corpus of any language, the acme of linguistic resources, is characterized by two fundamental features. First, it is essential that the corpus represent the entire language in question. This means that it should contain all types of communication, both written and oral, in all genres, from the belletristic to the dialectal, and represent all historical periods, from antiquity to the present. Second, it should be maximally balanced insofar as the text types in the corpus

correspond to their proportion of usage in real-life communication to the extent that it is feasible, taking into consideration aspects such as data availability and legal restrictions.

A national corpus makes it possible to conduct a wide range of linguistic analyses into the language for which it is available. As the creators of the Russian National Corpus (hereafter RNC) explain:

> [Electronic] libraries are not well suited to academic work on the nature of language; they tend to focus on the content of texts rather than their language properties, while the creators of the Corpus recognize the importance of literary or scientific value of the texts, but see them as a secondary feature. Unlike an electronic library, the National Corpus is not a collection of texts which are deemed "interesting" or "useful" of themselves; the texts in the Corpus are interesting and useful for the study of language. Such texts might include not only great works of literature, but also works of a "secondary" writer, or a transcription of an ordinary conversation. (http://www.ruscorpora.ru/en/corpora-intro.html)

Since the RNC became available in 2004, it has developed into a functional and extensively annotated resource. Today, in terms of its size and scientific value, it is comparable to the American, British, Czech, and Polish national corpora. The core collection of the RNC includes manually selected samples of written and spoken texts. Those samples represent various genres, such as fiction, drama, memoirs, news and literary criticism, popular non-fiction and textbooks, religious and technical texts, business and jurisprudence papers, and texts on daily life. The samples include texts that were not initially intended for publication.

Any national corpus by definition is large and multifaceted. At the time of writing, all subcorpora and spin-off projects available on the ruscorpora.ru site comprise 600 million tokens. Table 17.1 lists the detailed statistics on the main

Table 17.1 Russian National Corpus: texts by subcorpora

Subcorpora	*Number of texts*	*Number of sentences*	*Number of tokens*	*% of tokens*
The main subcorpus	76,882	17,574,752	209,198,275	57.3
The news-media subcorpus	181,175	8,553,495	113,292,003	31.0
The dialectal subcorpus	197	20,273	194,283	0.1
The educational subcorpus	229	65,666	664,751	0.2
The parallel subcorpus	370	1,609,609	24,022,437	6.6
The poetry subcorpus	41,448	638,861	6,738,474	1.8
The oral subcorpus	3034	1,604,626	10,122,579	2.8
The multimodal subcorpus	31,741	148,619	648,576	0.2
In total:	*335,076*	*30,215,901*	*364,881,378*	*100*

Source: http://www.ruscorpora.ru/corpora-stat.html. The English translation is ours

Table 17.2 Russian National Corpus: texts by creation date (the main subcorpus only)

Periods	*Number of texts*	*Number of sentences*	*Number of tokens*	*% of tokens*
1701–1750	298	27,090	590,541	0.3
1751–1800	979	176,207	2,981,803	1.4
1801–1850	1098	704,678	10,380,375	4.8
1851–1900	2063	2,366,209	31,761,447	14.7
1901–1950	26,325	4,646,823	53,445,536	24.7
1951–2000	14,486	6,172,190	67,252,763	31.0
2001–2010	31,491	4,094,011	50,231,677	23.2
In total:	*76,740*	*18,187,208*	*216,644,142*	*100*

Source: http://www.ruscorpora.ru/corpora-stat.html. The English translation is ours

parts of the collections; Table 17.2 provides additional details on the core collection. The represented time periods vary due to the availability of the digitized sources of the particular period.

All the subcorpora are lemmatized, which occurs when all forms of a word are arranged under a headword as in dictionary form, called *lemma*, and annotated both morphologically and syntactically. Some of the subcorpora are also analyzed semantically (grouped in lexical classes according to the meaning) and derivationally (grouped by word formation). The crowning touches of this monumental resource are its diverse rich metadata and sophisticated search options, such as multiword expressions, tag repetition in adjacent tokens, and stress marking.

The site also hosts several spin-off projects of which the most interesting is the Old Russian subcorpus (Pichhadze 2005), which includes original Old Russian texts (such as chronicles and Novgorodian birch-bark letters) as well as translations from Greek texts (e.g., *The Romance of Alexander*, Flavius Josephus's *Books of the History of the Jewish War against the Romans*) and South Slavic texts, rewritten in Old Russian (e.g., *Izbornik* [Miscellany] of 1076). Other notable projects are the SynTagRus corpus (Boguslavsky et al. 2000), which is manually annotated with syntactic dependency and lexical function markups, and the FrameBank (Lyashevskaya and Kashkin 2015), which is annotated with semantic roles. To the best of our knowledge, the RNC is also the only resource that includes a corpus of Russian poetry, which allows searches by meter and rhyme of poetic texts from the eighteenth century to the present (Grishina et al. 2009).

17.4.1.1 Case Study: Tracking Collective Memory Through "Decade Constructions"[3]

The study of collective memory is a strong interdisciplinary field that concentrates on the exploration of collective mnemonic concepts. The aim is to analyze how and why people and society think about and collect the events of their mutual past. This research objective has drawn the attention of historians,

scholars of cultural studies, and anthropologists. However, this has almost never been addressed by linguists, despite the generally acknowledged importance of language as a key translator of culture (Lotman 2009; Koselek 2004). Attempts to explore the Russian collective memory through corpus analysis have been made by Bonch-Osmolovskaya (2018) and Götzelmann et al. (2019). The former analysis focuses on the constructions, which include a word-denoted decade preceded by an epithet, such as *lihie devânostye* (wild nineties), *zolotye pâtidesâtye* (golden fifties), and *groznye tridcatye* (terrible thirties). We refer to them hereafter as *decade constructions*. The basic assumption is that these constructions reflect the mnemonic patterns of each decade in Soviet and post-Soviet history; hence, their linguistic analysis makes it possible to reconstruct patterns of collective memory.

The data obtained from the Russian National Corpus have been re-organized so that the final dataset had a total of 242 sentences with decade constructions, which refer to the period from the 1920s until the 1990s. A non-trivial semantic feature of this construction is that the ordinal, such as *dvadcatye* (twenties), refers to a timespan that does not fully coincide with a corresponding decade. A timespan is perceived as a featured historical period, with specific connotations, expressed by an adjective and shared between a speaker and an audience. As Zerubavel (2003, 31) observes, the corpus analysis of decade constructions reveals a non-even distribution of historical periods so that "hills and valleys" appear in the collective memory. Some decades seem to be salient and prominent mnemonic concepts, whereas others remain almost forgotten.

Frequency analyses of the examples have their own methodological specificity. Most corpus methods focus on the most frequent entries, and those that are statistically non-significant are typically not considered. In this case, however, even a unique entry should not be neglected and must be included in the analysis, as the adjective still refers to a shared collective concept that can only be understood if this association occurs. Figure 17.1 displays the overall frequency distribution of the construction for each decade. The radar-chart values for each decade correspond to the mean value for all constructions. Table 17.3 presents the number of constructions that occur in the RNC for each ordinal.

It is clear from both Fig. 17.1 and Table 17.3 that the distribution is not even. Naïve chronology covers almost all of the decades in the twentieth century, but some are more important (the 1930s and 1990s). Some decades are rarely referred to (the 1950s and 1980s), which means that they do not form a mnemonic pattern and barely exist in the collective memory. The 1990s, which was the turbulent period of post-Soviet political and economic transition and a time of intensive and highly emotional social reflection, display the highest frequency, whereas the 1950s and the 1980s represent the lowest, which is less than the overall mean. These two periods coincide with the end of two historical epochs: Stalin's reign of terror and Brezhnev's era of stagnation. One might speculate that they do not form a holistic mnemonic pattern because they are more likely to represent a rupture between the preceding and subsequent decades.

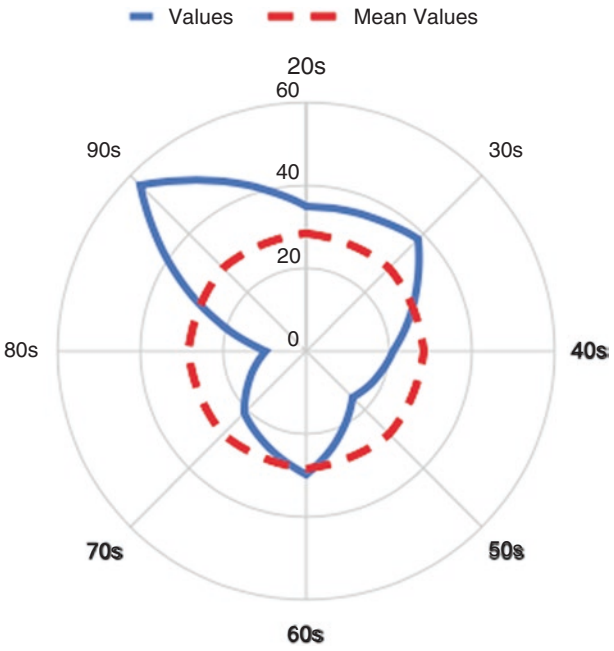

Fig. 17.1 Frequency of adjective decade constructions for each decade

Table 17.3 Frequency of adjective decade constructions for each decade

Decades	*Construction frequency*
1920s	35
1930s	38
1940s	21
1950s	16
1960s	30
1970s	21
1980s	10
1990s	57

Our research continues with the analysis of 117 adjectives, which are used with the ordinals in question and fall into several semantic classes. The first three classes are united by the semantics of direct or indirect emotional assessment toward an ordinal. The epithet basically defines the decade as a separate cultural phenomenon with specific symbolic meaning; the epithet also contains a built-in assessment of the epoch by the speakers. These are adjectives that refer to real-world attributes that are characteristic of the historical period, such as *ateističeskie dvadcatye* (atheistic twenties), *stilâžnye pâtidesâtye* (dandy fifties), and *banditskie devânostye* (gangster nineties). Another major class

comprises adjectives of positive or negative assessment, which include metaphorical expressions, such as *lihie devânostye* (wild nineties). There are also adjectives that emphasize the prominence of the decade that cannot be classified as either positive or negative, such as *nepovtorimye devânostye* (unique nineties) and *rokovye sorokovye* (fatal forties). Two more adjectival classes are connected by spatial or geographical references, such as *sovetskie semidesâtye* (Soviet seventies) and *moskovskie šestidesâtye* (Moscow sixties), or by temporal references of which the most frequent are *rannie* (early) and *pozdnie* (late). One might expect the latter two to reflect a common characteristic of any decade, but this is not the case because their distribution across the decades is uneven (see Fig. 17.2): the concept "early/late" is not selected randomly but corresponds to micro-historical patterns. Hence, the "early thirties" is a period that precedes the Great Terror, which is not referred to as the "late thirties" because it has its own name. On the other hand, the "late fifties" and "early sixties" combined constitute the conceptual memory of *the Khrushchev Thaw* (Rus. *ottepel'*).

As Fig. 17.1 indicates, "the nineties" is the most frequently occurring nomination in the dataset and it represents a very special case of collective memory modeling. Approximately 70 percent of all "nineties" examples contain attributes of either a positive or negative assessment. The most common is *lihie devânostye* (wild nineties), which occurs 14 times (30%). However, on 10 of those occasions, the adjective *lihie* (wild) is enclosed within quotation marks, which makes the whole pattern more complex. One might assume that the speaker uses quotation marks to refer not to the collective memory but to the

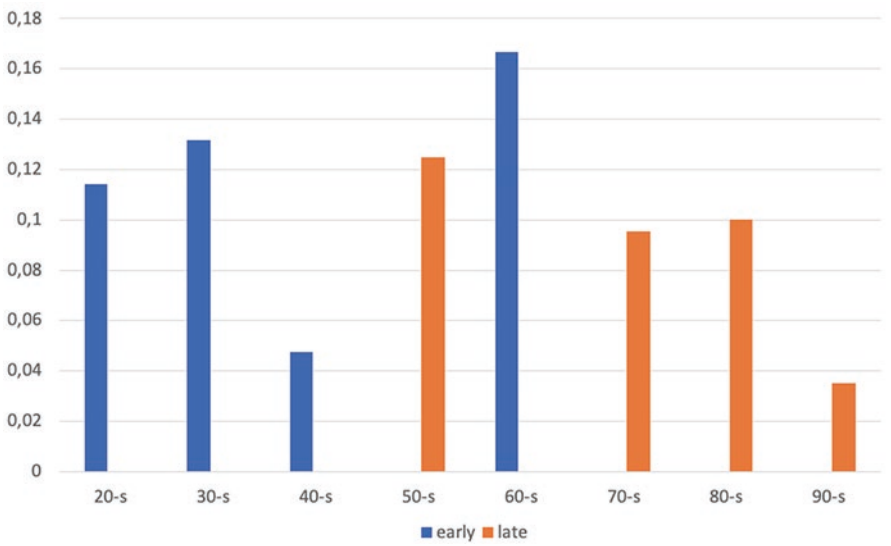

Fig. 17.2 Distribution of *rannie* (early) and *pozdnie* (late) in decade constructions

preceding contextual usage of the expression specifically adopted by those in power. This is where the process of lexicalization begins and initial conceptual semantics fade. This is even more obvious when examining the Newspaper subcorpus within the RNC (about 133 million tokens from 2001 to 2014). The "nineties" constructions again constitute the dominant majority, comprising approximately 50 percent of all the examples, of which 30 percent is *lihie devânostye* (wild nineties). However, the marked difference in distribution demonstrates that the collective memory of the post-Soviet nineties was formed later in the noughties when it became a phrasal cliché through the perpetual repetition of *lihie devânostye* (wild nineties) in the media. Figure 17.3 presents the rapidly increasing frequency of the "wild" nineties compared to all other adjectives followed by the ordinal; "wild" becomes nearly dominant from 2008 to present. Having become a fixed-word combination, *lihie devânostye* (wild nineties) no longer triggers collective memory but is instead a meme, a semantically bleached language sign that has nothing in common with the concept of "wildness and chaos," which is something that could be associated with the period in question.

The case study presented above demonstrates the potential usefulness of relatively small datasets in collecting promising historical observations on "memory landscapes" by using linguistic corpora. Although the dataset is too small to apply standard statistical measures, the qualitative analysis of symbolic value provides an alternative basis for interpretation, which is based on evidence rather than statistics. There is no single occurrence of the construction nor is a single use of the adjectives random because they are all bricks in the construction of a controversial and multifaceted collective memory. What is of significance here is the reliability of the data: it is a corpus that is balanced in

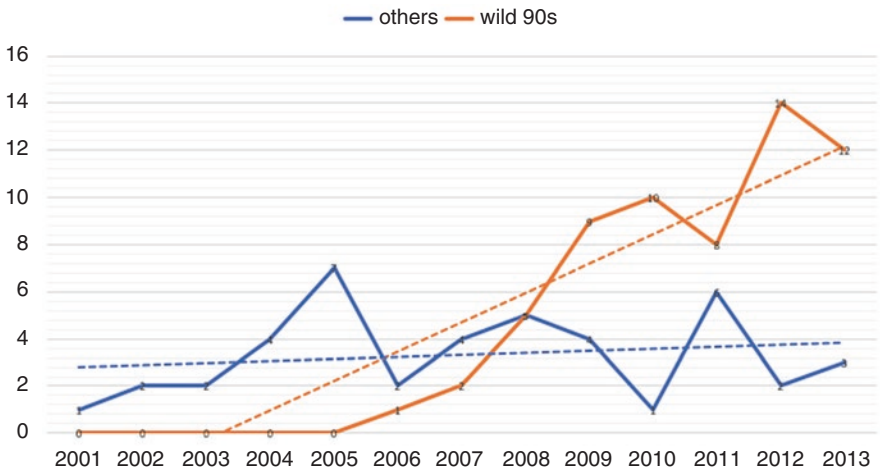

Fig. 17.3 Frequencies of *lihie devânostye* (wild nineties) compared to all adjectives attested in the construction (2001–2013, the Newspaper subcorpus)

terms of both genre and timespan. Its morphological mark-up also allows the user to search not only for a word but also for a moving context that yields insights that are otherwise inaccessible.

17.4.2 Integrum (www.integrumworld.com)

Although it is not a corpus in the strict sense, the Integrum database of Russian media has features that render it extremely useful for research purposes in comparison with both linguistic corpora and biased raw Internet data (for a comparison, see Mustajoki 2006; Plungian 2006). The service is not free, but libraries and universities throughout the world provide access to it online.

The main benefit of Integrum is that it covers almost all newspapers and magazines published in Russia from the beginning of the 1990s. Thus, users have easy access to the full texts of metropolitan media publications, such as *Izvestia* and *Komsomolskaya pravda,* as well as to far more remote and thus difficult to obtain media including *Vesti respubliki* (Grozny, Chechnya), *Večernij Murmansk*, and *Saratovskaya panorama*. Dozens of Russian-language newspapers published outside Russia are likewise available, including *Evropa-Èkspress* (Berlin), *Karavan* (Kazakhstan), and *Minskij kur'er* (Belarus). Complementing the printed media, Integrum also includes a wide variety of data from radio and television broadcasts, online media, news agencies, and legislation. A total of approximately 200 million texts are available, which means many more than 50 billion running words.

A researcher can find some of the materials available in Integrum elsewhere on the Internet. Yet what makes Integrum invaluable is the thorough categorization of the data. Within the categories, users can search for further sources of interest simply by clicking on a given list of resources. This option is especially useful for those who are interested in examining different opinions on political issues, such as pension reforms throughout Russia, or in comparing regional differences in attitudes, such as how foreign powers are perceived in the eastern part of Siberia versus attitudes that prevail in the capital region.

The data in Integrum are not deeply morphologically annotated, but the search options are diverse nonetheless. To make searches, users can utilize tokens (word forms), lemmas (words), or parts of words (using wildcards). It is possible to determine the distance between the searched words, that is, how far apart they are to be included in the results. For example, the query [*modernizac** :3 *Rossi**] returns all contexts in which all forms of the words occur within one to three words of each other. In addition, a brief excerpt and the full text are provided for the examples found. Researchers may also conduct more sophisticated searches to create macros that enable them to more precisely pinpoint the passages they find most interesting and useful. For anyone with a limited command of Russian, one available option is to make a quick automatic English translation in the search box. Thus, a look-up value, such as "digital Russia," returns texts containing corresponding Russian words highlighted in Russian-language articles.

17.4.2.1 Case Study: Political Buzzwords in Russian[4]

Integrum is intended primarily for business people, journalists, and scholars who are interested in Russian society and politics, and the economy, but it can also be used effectively in linguistic studies, such as to determine how people use the Russian language (see Mustajoki and Pussinen 2006, 2008). Below we present a case that demonstrates the use of Integrum in interdisciplinary research to examine attitudes toward the modernization process of the Russian media, with special reference to events that made modernization impossible to achieve.

Although "modernization," or *modernizaciâ* in Russian, has a colloquial usage, its appropriation by Dmitry Medvedev's administration made it a buzzword that is identifiable as a marker in certain types of political discourse. This word became a central concept in Medvedev's political program during his presidential term (2008–2012). Thus, *modernizaciâ* has both political and economic connotations and has continued to be associated with Medvedev and his politics.

In their study of media texts on modernization, Laine and Mustajoki (2017) concentrated on the period from December 31, 2000, to December 31, 2012, because it covers the rise and fall in usage of this notion in Russian media discourse. Within that timeframe, 94,500 occurrences of the word *modernizaciâ* in all its forms were detected in 350 national Russian newspapers (see Fig. 17.4).

A preliminary investigation of the examples revealed that discussion related to the concept was frequent, but that the overall attitude was rather skeptical.

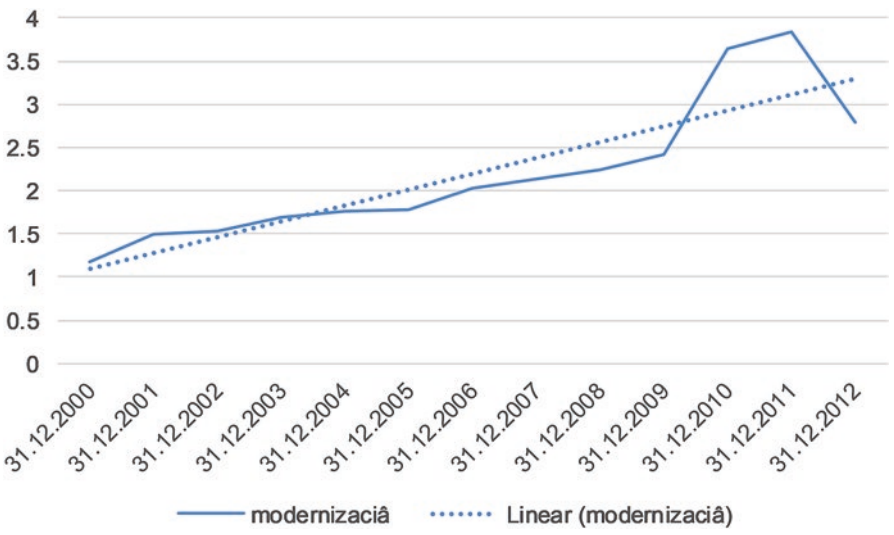

Fig. 17.4 The relative frequency (%) of *modernizaciâ* (modernization) occurring in texts from Russian national newspapers (Source: Integrum, Dec. 31, 2000—Dec. 31, 2012)

Many writers welcomed the modernization process per se, but they expected it to fail as did all previous attempts at reform. They insisted that reform could only succeed if *X* were to take place, *X* being something specific that should be undertaken, as in the following example:

> *Bez ètoj mobil'nosti nevozmožna modernizaciâ strany, a značit, gosudarstvu pridetsâ pojti na strukturnye izmeneniâ v èlitah.* (RBC, July 1, 2008)
>
> Without this mobility, modernization of the country is impossible, which means that the state will have to go for structural changes in the elites. (RBC, July 1, 2008)

Our observation corresponds to that of Juri Prokhorov and Iosif Sternin (2006, 67–68), who claimed that Russians tend to typically adopt reasoning based on a "single-explanation" in their public representations of themselves. These sociolinguists examined the cliché that Russians search for a centralized solution for all problems and put their trust in quick and simple resolutions for complex problems. According to Prokhorov and Sternin, what lies behind this stereotype is the historically grounded, left-leaning reasoning that responsibility for everything rests with *oni* (in Russian, "they"; here, "the ones with power"). This responsibility pertains to not only the country's prosperity but also the well-being of the nation. "They" may be personalized, as a czar or a president, or it may be an abstract concept referring to those who have power. The implicit belief underlying this attitude is that the solution lies outside and above, not with the people themselves, whereas "they"—the ones with the power—have the opportunity, the capability, to make life better in Russia.

Laine and Mustajoki (2017) used the multistage cascade search technique to explore that line of argument more deeply as it applies to the concept of modernization. As a first step, all contexts of all forms of the word *modernizaciâ* were extracted. Thereafter, only contexts that referred to the modernization of the whole country were considered further, rather than those that related to a specific sector, such as transportation, education, or the army. To achieve this, they introduced additional search criteria: contextual conditions, which restrict the context to all-Russian modernization, for example, *modernizaciâ + Rossii* (modernization of Russia) or *modernizaciâ strany* (modernization of the country). More detailed restrictions were applied during the next step—finding the "single-explanation" argument. This means that certain expressions had to be attested in a nearby context within the same sentence, such as [*modernizaciâ*] *vozmožna, tol'ko esli* ([modernization] is possible only if) or [*dlja modernizacii*] *neobhodimo* ([for modernization,] it is necessary to). The corpus was restricted to the news media, which excluded scientific articles, official documents, and historical texts. In total, approximately 100 contexts were subject to further detailed analysis.

To summarize, according to the results by Laine and Mustajoki, the factors that obstruct modernization fall into several categories: (a) economic (such as a low level of investment in industry, raw-material dependency and a lack of

"civilized" competition); (b) scientific and educational (the country should create the "necessary" environment for young scientists and "normal" conditions for specialist education in order to avoid a national brain drain); and (c) political (controversial opinions such as "Under this rule, modernization is impossible" and "Only Putin would have the ability to modernize the country"; "The party in power, United Russia, can ensure the success of modernization").

The Russian word *importozamêsenie*, which is both difficult to pronounce and comprehend, means "import substitution" in a Russian-specific sense. A new phase of Russian political rhetoric began in 2012 when Putin embarked on his successive terms in the Kremlin. The context of both his third and fourth terms was that of empowered authoritarianism. After the annexation of Crimea, the European Union (EU), the United States (US) and some other countries imposed sanctions on Russia, and Russia enacted counter-sanctions on EU products (see Travin et al. 2020). In the changed political situation, President Putin introduced the new concept of *importozamêsenie* (import substitution), among other buzzwords. Its meteoric rise in the media is astonishing and comparable with that of "Russian modernization"; Fig. 17.5 illustrates how quickly its frequency increased in Russian media coverage from 2014 onward.

The "single-explanation" comments were again attested in the data after the new buzzword appeared. This time, the explanations tempering the effect of *importozamêsenie* (import substitution) included the competitiveness of Russian enterprises and a new attitude toward agriculture:

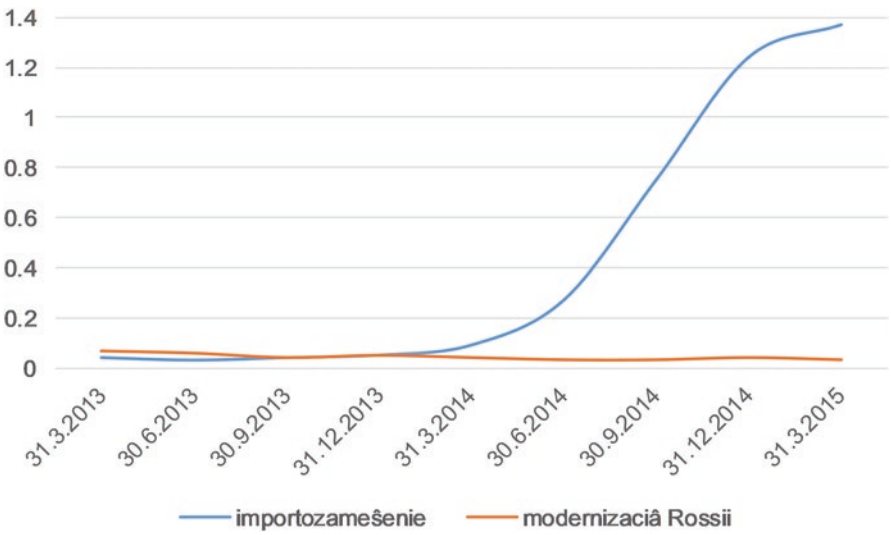

Fig. 17.5 The usage of *modernizaciâ Rossii* (modernization of Russia) in comparison to *importozamêsenie* (import substitution) (Source: Integrum, Russian National Media, 2013–2015)

Vozmožno, ambicioznye plany èkspertov sel'hozotrasli po importozamêseniû i sbudutsâ, no tol'ko esli rynok tepličnyh ovôšej budet horošo udobren bankovskimi investiciâmi i gosudarstvennoj podderžkoj. (*Rossiyskaya gazeta*, September 4, 2015)

Perhaps the ambitious plans of experts in the agricultural sector for import substitution will come true, but only if the greenhouse vegetables market is well-fertilized with bank investments and government support.

[S]trane nužno importozamêsenie, no ono vozmožno tol'ko pri nizkoj inflâcii. (*Sovetskaya Rossiya*, November 20, 2014)

[T]he country needs import substitution, but it is possible only with low inflation.

To summarize, the large-scale media data provided by Integrum revealed three major findings. First, a large amount of data distinctly reflect the extent to which awareness of the political agenda set by Russian leaders is spreading among people. The concepts of "modernization" and "import substitution" aroused interest, having been introduced by leaders and reproduced in the media. Second, a more detailed analysis revealed recurring attitudes toward the concepts: there were frequent occurrences of "single-explanation" reasoning concerning the possibilities of modernization and import substitution, which appears to be a recurrent argument in Russian media discourse. Third, a qualitative analysis made it possible to identify the reasons that were used in media discourse to prevent changes in Russia. A single reason was usually provided to explain the failure, be it economic, educational, or political.

17.5 Conclusion

Texts are the principle sources of analysis in various types of research. Large textual corpora are an excellent source for investigating diverse concepts and their reflection in the language and attitudes in a society. These types of studies need both statistical data and in-depth analysis, which the described resources have to offer. If a researcher is aware of how to use the available resources and conducts an investigation within the limits that the data impose, then the results are reliable and inspiring.

We have presented various textual resources that are available for Russian studies: the web as a corpus, electronic libraries, and linguistics corpora. Some of these are specifically designed for linguistic research, but the majority may be effectively utilized in wider text-based studies. We emphasized the two most significant resources in particular: the Russian National Corpus and the Integrum database. The case studies we presented utilized a basic corpus-informed analysis to illustrate the usefulness of both resources in the study of societal changes as they are reflected in the language.

Notes

1. The projects embrace the cause of promoting copyleft ideas, the free distribution of copies (https://en.wikipedia.org/wiki/Copyleft). Although many of the publications on these sites are no longer under copyright, there have been many accusations of copyright infringement.
2. This section is adapted from our previous review by Kopotev et al. (2018). Readers who are interested in the specific linguistic details are advised to consult that publication.
3. This section is based on Bonch-Osmolovskaya (2018), where more details are provided.
4. This section is based to some extent on Laine and Mustajoki (2017), where more details are provided.

References

Belikov, Vladimir, Alexander Piperski, Vladimir Selegey, and Serge Sharoff. 2013. Big and Diverse Is Beautiful: A Large Corpus of Russian to Study Linguistic Variation. In *Proceedings of the 8th Web as Corpus Workshop (WAC-8)/International Conference on Corpus Linguistics*, Lancaster.

Benko, Vladimír. 2014. Aranea: Yet Another Family of (Comparable) Web Corpora. In *International Conference on Text, Speech, and Dialogue*, 247–256. Cham: Springer.

Boguslavsky, Igor, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Kreidlin, and Nadezhda Frid. 2000. Dependency Treebank for Russian: Concept, Tools, Types of Information. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING), 987–991*. Saarbrücken.

Bonch-Osmolovskaya, Anastasia. 2018. Imena vremeni: èpitety desâtiletij v Nacional'nom korpuse russkogo âzyka kak proekciâ kul'turnoj pamâti [Names of Time: Epithets of Decades in the Russian National Corpus as a Projection of Cultural Memory]. *Shagi/Steps* 4 (3): 115–146.

Bozdag, Engin. 2013. Bias in Algorithmic Filtering and Personalization. *Ethics and Information Technology* 15 (3): 209–227. https://doi.org/10.1007/s10676-013-9321-6.

Dobrushina, Nina, ed. 2007. *Nacional'nyj korpus russkogo âzyka i problemy gumanitarnogo obrazovaniâ* [The Russian National Corpus and Issues in Humanitarian Education]. Moscow: Higher School of Economics.

Erjavec, Tomaž, Ivan Deržanski, Dagmar Divjak, Anna Feldman, Mikhail Kopotev, Natalia Kotsyba, Cvetana Krstev, et al. 2010. *MULTEXT-East Non-commercial Lexicons 4.0*. Slovenian Language Resource Repository CLARIN.SI. Accessed June 1, 2019. http://hdl.handle.net/11356/1042.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80: 298–320. Accessed October 9, 2017. https://doi.org/10.1093/poq/nfw006.

Götzelmann, Michael, Kirill Postoutenko, Olga Sabelfeld, and Willibald Steinmetz. 2019. The Historical Semantics of Temporal Comparisons through the Lens of Digital Humanities: Promises and Pitfalls. In print. Accessed December 15, 2019. https://www.academia.edu/41122839/The_Historical_Semantics_of_Temporal_Comparisons_through_the_Lens_of_Digital_Humanities_Promises_and_Pitfalls.

Grishina, E.A., K.M. Korchagin, V.A. Plungyan, and D.V. Sichinava. 2009. *Poètičeskij korpus v ramkah NKRÂ: obŝaâ struktura i perspektivy ispol'zovaniâ* [The Poetic Corpus in RNC: Its Structure and Prospects of Use]. In Nacional'nyj korpus russkogo âzyka: 2006–2008 [Russian National corpus: 2006—2008], ed. V.A. Plungyan, 71–113. Saint-Petesburg: Nestor-istoriâ.

Jakubíček M. et al. 2013. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL*. 125–127.

Kilgarriff, Adam. 2001. Web as Corpus. *Proceedings of the Corpus Linguistics Conference (CL 2001). University Centre for Computer Research on Language Technical Paper Vol. 13, Special Issue, Lancaster University*, 342–344. http://ucrel.lancs.ac.uk/publications/CL2003/CL2001%20conference/papers/kilgarri.pdf.

Kopotev, Mikhail, Olga Lyashevskaya, and Arto Mustajoki. 2018. Russian Challenges for Quantitative Research. In *Quantitative Approaches to the Russian Language*, ed. Mikhail Kopotev, Olga Lyashevskaya, and Arto Mustajoki, 3–29. Abingdon: Routledge.

Koselek, R. 2004. *Futures Past: On the Semantics of Historical Time. Series: Studies in Contemporary German Social Thought*. New York: Columbia University Press.

Laine, Veera, and Arto Mustajoki. 2017. Preconditions for Russian Modernisation: A Media Analysis. In *Philosophical and Cultural Interpretations of Russian Modernisation*, ed. Katja Lehtisaari and Arto Mustajoki, 175–190. Abingdon: Routledge.

Lotman, Yu. 2009. *Culture and Explosion* (*Semiotics, Communication and Cognition*). Translated by Wilma Clark and edited by Marina Grishakova. De Gruyter Mouton.

Lyashevskaya, O. 2016. *Korpusnye instrumenty v grammatičeskih issledovaniâh russkogo âzyka* [Corpus Tools in Grammatical Studies of the Russian Language]. Moscow: LRC Publishing House.

Lyashevskaya, O., and E. Kashkin. 2015. FrameBank: A Database of Russian Lexical Constructions. In *International Conference on Analysis of Images, Social Networks and Texts*, 350–360. Cham: Springer.

McEnery, Tony, and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. Routledge.

Mitrenina, Olga. 2014. The Corpora of Old and Middle Russian Texts as an Advanced Tool for Exploring an Extinguished Language. *Scrinium* 10 (1): 455–461.

Mustajoki, Arto. 2006. The Integrum Database as a Powerful Tool in Research on Contemporary Russian. In *Integrum: točnye metody i gumanitarnye nauki*, ed. Galina Nikiporets-Takigava, 50–76. Moscow: Letnij sad.

Mustajoki, A., and O. Pussinen. 2006. Počemu narodu mnogo, ili Novye nablûdeniâ nad upotrebleniem vtorogo roditel'nogo padeža v sovremennom russkom âzyke [Why *narodu mnogo*: New Observations on the Use of the Second Genitive in Russian]. In *Integrum: točnye metody i gumanitarnye nauki*, ed. G. Nikiporets-Takigava, 50–75. Moscow: Letnij sad.

———. 2008. Ob èkspansii glagol'noj pristavki PO- v sovremennom russkom âzyke [Expansion of the Prefix PO in the Contemporary Russian]. In *Instrumentarij rusistiki: korpusnye podhody* (= Slavica Helsingiensia 34), 247–275. Helsinki.

Nivre, Joakim, Mitchell Abrams, Željko Agić. et al. 2018. *Universal Dependencies 2.3*, LINDAT/CLARIN Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2895.

Pichhadze, A.A. 2005. *Korpus drevnerusskih perevodov XI-XII vv. i izučenie perevodnoj knižnosti Drevnej Rusi* [The Corpus of Old-Russian Translations from 11–12 Centuries and Study of Translated Literature of Ancient Russia]. In *Nacional'nyj korpus russkogo âzyka: 2003–2005* [Russian National corpus: 2003—2005], ed. V.A. Plungian, 251–262. Moscow: Indrik.

Plungian, V.A. 2006. 'Integrum' i Nacional'nyj korpus russkogo âzyka v lingvističeskih issledovaniâh [Integrum and the Russian National Corpus in linguistic research]. In *Integrum: točnye metody i gumanitarnye nauki*, ed. G. Nikiporets-Takigava, 76–84. Moscow: Letnij sad.

———., ed. 2009. *Nacional'nyj korpus russkogo âzyka: 2006—2008* [Russian National Corpus: 2006—2008], 71–113. Saint-Petersburg: Nestor-istoriâ.

Plungian, V.A., and L. Shestakova. eds. 2014. *Korpusnyj analiz russkogo stiha* [Corpus Analysis of Russian Verse], 2. Moscow: Azbukovnik.

Prokhorov, Y.A., and I.A. Sternin. 2006. *Russkie: kommunikativnoe povedenie* [Russian Communication Strategies]. Moscow: Flinta.

Sharoff, Serge, and Joakim Nivre. 2011. The Proper Place of Men and Machines in Language Technology: Processing Russian Without any Linguistic Knowledge. In *Proceedings of Dialogue 2011, Russian Conference on Computational Linguistics.*

Shavrina, T., and O. Shapovalova. 2017. To the Methodology of Corpus Construction for Machine Learning: Taiga Syntax Tree Corpus and Parser. In *Proceedings of "CORPORA-2017" International Conference*, Saint-Petersburg, 78–84.

Travin, Dmitry, Vladimir Gel'man, and Otar Marganiya. 2020. *The Russian Path: Ideas, Interests, Institutions.* Stuttgart: ibidem Press.

Usage. 2019. *Usage of Content Languages for Websites.* Accessed November 28, 2019. https://w3techs.com/technologies/overview/content_language/all.

Zabotkina, V.I., ed. 2015. *Metody kognitivnogo analiza semantiki slova. Komp'úterno-korpusnyj podhod* [Methods in Cognitive Analysis of Word Semantics: A Computational and Corpus-based Approach]. Moscow: Yazyki Slavyanskoi Kultury.

Zerubavel, E. 2003. *Time Maps. Collective Memory and the Social Shape of the Past.* Chicago: The University of Chicago Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.